# REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188



| 1. REPORT DATE (DD-MM-YYYY) | 2. REPORT TYPE | 3. DATES COVERED (From - To) |
|---|---|---|
| | Technical Report | - |

**4. TITLE AND SUBTITLE**
IU Progress Report May 2013

**5a. CONTRACT NUMBER**
W911NF-12-1-0037

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHORS**
Filippo Menczer, Alessandro Flammini, Qiaozhu Mei, Sergey Malinchik

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAMES AND ADDRESSES**
Indiana University at Bloomington
Trustees of Indiana University
509 E 3RD ST
Bloomington, IN 47401 -3654

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

**10. SPONSOR/MONITOR'S ACRONYM(S)**
ARO

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**
61766-NS-DRP.20

**12. DISTRIBUTION AVAILIBILITY STATEMENT**
Approved for public release; distribution is unlimited.

**13. SUPPLEMENTARY NOTES**
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

**14. ABSTRACT**
The team worked at the evaluation of its algorithm to discriminate between trending and promoted topic on Twitter, with excellent results.
We continued developing the underlying SAX-VSM approach to represent time series as its proving a convenient and effective framework to all sorts of classification effort.

**15. SUBJECT TERMS**
Promoted/trending topics, time series classification, SAX algorithm

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 15. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | Alessandro Flammini |
| UU | UU | UU | UU | | 19b. TELEPHONE NUMBER 812-856-1830 |

**Report Title**

IU Progress Report May 2013

**ABSTRACT**

The team worked at the evaluation of its algorithm to discriminate between trending and promoted topic on Twitter, with excellent results.

We continued developing the underlying SAX-VSM approach to represent time series as its proving a convenient and effective framework to all sorts of classification effort.

**DARPA SMISC Project:**
DESPIC: Detecting Early Signatures of Persuasion in Information Cascades

**Teams:**
Indiana University: A. Flammini (PI) and F. Menczer
University of Michigan: Qiaozhu Mei
Lockheed Martin Advanced Technology Laboratories (ATL): S. Malinchik

## Progress Report – May 2013

**IU:** During the month of May 2013 the Indiana team worked on the evaluation of the trending topics and promoted content classification system.

The promoted content detection system is fully operative and it is collecting the trends in real-time from the Twitter platform and classifying those that are genuinely trending versus engineered ones. The system collected a dataset including over 100 promoted trends and over 20 thousands genuine trends, and uses a combination of machine-learning and information theory techniques to generate features from the multivariate time-series extracted from the data to the purpose of classify them. As of the date, we can generate a set of over 200 features that include: 1) network features of the diffusion network of each given trending meme (e.g., density, degree distribution, etc.); 2) temporal features (e.g., intervals between retweets, mentions, etc.); 3) user features (including followers number, statistics on users' behaviors etc.)

Once the time-series of these features are generated, the system builds a symbolic representation of such data by using a novel technique knows as SAX (Symbolic Aggregate approXimation) [1] to encode the time-series allowing for dimensionality-reduction and data compression. Each time-series is built so that to consist of a sequence of 432 data-points, each data-point representing the value of the given features in a 20min length. The SAX encoding first takes the time-series and splits it in 400 overlapping subsequences consisting of 32 data-points each, starting from the first data-point and shifting of one position each time. The SAX algorithm processes each subsequence of the time-series, further dividing that interval in 5 chunks; each chunk is then represented as a 5-word-long SAX sequence, by using a 5-letter-long alphabet.

The encoding of time-series by using SAX allows us to incorporate in the data the temporal dimension that would be otherwise disregarded if we adopted data-points as independent by dealing with the raw time-series using a standard machine-learning classification approach.

The classification system is trained to learn classes from these SAX-encoded time-series of features and adopts the following algorithms to classify content: ensemble methods with voting (e.g., Random Forest, ExtraTree) and boosting (e.g., AdaBoost), Support Vector Machines (SVM) and Hidden Markov Models (HMM).

In the course of our preliminary evaluation we observed that the SAX-encoding pre-processing step that incorporate temporal dependence of data-points allows the system to outperform the previous version that did not adopt the SAX-encoding process. In Table 1 we report a comparison between the performance of the two different approaches in the promoted content versus genuine trends classification task by using ExtraTree, Random Forest and SVM. In particular, this analysis highlights that the system achieves accuracy close to 90% both by using SVM and Random Forest.

*Table 1: Performance of different classifiers with and without the SAX encoding.*

| Classifier | Without SAX | SAX (all features) | SAX (top 10 features) |
|---|---|---|---|
| **ExtraTree** | 0.644 | 0.828 | 0.860 |
| **Random Forest** | 0.644 | 0.850 | 0.893 |
| **SVM** | 0.555 | 0.746 | 0.894 |

## *References*

[1] Patel, Pranav, et al. "Mining motifs in massive time series databases." Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. IEEE, 2002.

**UM:** PhD student Zhe Zhao presented our paper titled "Questions about questions: an empirical analysis of information needs on Twitter" at the 22nd international conference on World Wide Web (WWW'13), which is the top academic conference of research of the Web. The work presented in the paper was done under the support of this grant. The paper presents a set of techniques we developed to extract and analyze the information needs of Twitter users. Specifically, we first designed a classifier to identify real questions, in which the user expects an informative answer. Based on the classification results, we then analyze the dynamics of information needs within a longitudinal sample comprised of over a billion tweets. This work has meaningful impact in showing that both the volume and the entropy of information needs in Twitter react sensitively to real world events, and thus can be used as sensors in detecting such events and as predictors of search engine queries. The presentation is well received at the conference, with many positive feedbacks especially from researchers in search engine industry (e.g., Microsoft, Google).

**ATL:** In May 2013 ATL team continued working in parallel with Indiana team on classification problem of two types of activity on Twitter: advertisement campaigns defined as promoted content on Twitter and non-promoted naturally trending topics (see IU report). At this time we focused on exploring a new data set consisting

of 76 promoted and 853 non-promoted examples provided by Indiana team. Each temporal time series is characterized by 224 features (see Indiana's team part of this report). For accuracy evaluation of our classification approach based on SAX-VSM technology we used a well-known LOOCV (Leave-One-Out Cross-Validation) analysis. We also implemented a simple Monte Carlo algorithm to perform a search of the most discriminative cascade features. Our preliminary results indicate that with 16 best features we achieve a classification accuracy of **89.4**%.